AD NUMBER

ADB288332

NEW LIMITATION CHANGE

TO

Approved for public release, distribution unlimited

FROM

Distribution: Further dissemination only as directed by Office of Naval Research, 800 N. Quincy St., Arlington, VA 22217-5660, Apr 2003 or higher DoD authority.

AUTHORITY

ONR, per DTIC Form 55, dtd 24 Oct 2003

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT DATE	3. DATES COVERED *(From - To)*
15-04-2003	FINAL TECHNICAL REPORT	15 DEC. 2001- 31 DEC. 2002

4. TITLE AND SUBTITLE

AN EXPERIMENTAL EVALUATION OF TUTORIALS IN PROBLEM SOLVING (TIPS): A REMEDIAL MATHEMATICS TUTOR

5a. CONTRACT NUMBER

5b. GRANT NUMBER

N00014-02-1-0191

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

ROBERT K. ATKINSON
ARIZONA STATE UNIVERSITY

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

MISSISSIPPI STATE UNIVERSITY
BOX 6156
MISSISSIPPI STATE, MS 39762

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

OFFICE OF NAVAL RESEARCH
BALLSTON CENTRE TOWER ONE
800 NORTH QUINCY STREET
ARLINGTON, VA 22217-5660

10. SPONSOR/MONITOR'S ACRONYM(S)

ONR

11. SPONSORING/MONITORING AGENCY REPORT NUMBER

12. DISTRIBUTION AVAILABILITY STATEMENT

DISTRIBUTION STATEMENT F:
Further dissemination only as directed by
(APR 2003)
or higher DoD authority.

1

20030506 017

14. ABSTRACT

Our laboratory conducted an experimental evaluation of Tutorials in Problem Solving (TiPS), a computer environment representing a schema-based approach to training arithmetic and problem solving skills in remedial adult populations. Specifically, this project was designed to accomplish several objectives: (1) document the level of instructional adaptation provided by the TiPS system; (2) provide a basic evaluation of the overall TiPS system; (3) determine the amount of instructional time involved with the use of TiPS system; and (4) determine the affective nature of the TiPS learning experience. According to our evaluation project: (1) TiPS does monitor the learners' performance and adapt instructional delivery to meet their needs; (2) the average posttest score for participants in a TiPS group was significantly higher than their peers in a untreated control group; (3) the average time for completing the Tips computer instruction was 3.84 hours (SD = 0.97) and ranged from 1.83 to 6 hours; and (4) the learners exposed to TiPS reported feeling that the instructional material, including the examples and problems, helped them to understand how to approach and solve word problems and that, overall, the instructional environment was well designed.

15. SUBJECT TERMS

INTELLIGENT TUTORING SYSTEMS, REMEDIAL MATHEMATICS, WORKED EXAMPLES

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			19b. TELEPONE NUMBER *(Include area code)*

FINAL REPORT

Award
ONR N00014-02-1-0191

Title
An Experimental Evaluation of Tutorials in Problem Solving (TiPS): A Remedial Mathematics Tutor

Original Principle Investigator and Author of Report
Robert K. Atkinson,

Institution
Arizona State University

Date of Submission
April 15, 2003

Current Principle Investigator
Thomas Hosie

Institution
Mississippi State University

Table of Contents

ABSTRACT

Our laboratory conducted an experimental evaluation of Tutorials in Problem Solving (TiPS), a computer environment representing a schema-based approach to training arithmetic and problem solving skills in remedial adult populations. Specifically, this project was designed to accomplish several objectives: (1) document the level of instructional adaptation provided by the TiPS system; (2) provide a basic evaluation of the overall TiPS system; (3) determine the amount of instructional time involved with the use of TiPS system; and (4) determine the affective nature of the TiPS learning experience. According to our evaluation project: (1) TiPS does monitor the learners' performance and adapt instructional delivery to meet their needs; (2) the average posttest score for participants in a TiPS group was significantly higher than their peers in a untreated control group; (3) the average time for completing the Tips computer instruction was 3.84 hours (SD = 0.97) and ranged from 1.83 to 6 hours (excluding test-taking time); and (4) the learners exposed to TiPS reported feeling that the instructional material, including the examples and problems, helped them to understand how to approach and solve word problems and that, overall, the instructional environment was well designed.

SUMMARY

Objectives

The purpose of this project was to conduct an experimental evaluation of TiPS (Tutorials in Problem Solving), a computer environment representing a schema-based approach (e.g., Marshall, 1995) to training arithmetic and problem solving skills in remedial adult populations. Specifically, this project was designed to accomplish several objectives: (1) document the level of instructional adaptation provided by the TiPS system, (2) provide a basic evaluation of the overall TiPS system (3) determine the amount of instructional time involved with the use of TiPS system, and (4) determine the affective nature of the TiPS learning experience.

Approach

To accomplish these objectives, this project involved two phases in which rigorous empirical standards were applied during each phase. The initial phase, which focused on addressing the first objective, involved an analytic investigation of how the TiPS system behaves in response to various types of simulated performance. This analytic investigation was based on iterative "user" trials where graduate students in our lab used the TiPS system while adopting the behavior of "users" with a wide-range of ability levels. During their interaction with the system, its performance was systematically observed, which permitted us to examine whether the system behaved sensibly in terms of the ability level of the "user."

The second phase of our project addressed the remaining three objectives through the employment of a regression-discontinuity (RD) design, a pretest-posttest program-comparison group strategy where participants are assigned to program or comparison groups solely on the basis of a cutoff score on a pre-program measure (i.e., pretest). A basic type of RD design requires a preprogram measure (e.g., pretest), a post-program measure (e.g., posttest), and a measure that describes the assignment status of persons (received program or did not receive the program). The RD design is distinguished from the other pretest-posttest designs by its assignment strategy (Cook & Campbell, 1979). Basically, all participants are assigned to a program or comparison group in the basis of a cutoff score on the preprogram measure. Participants scoring on one side are assigned to the program while participants scoring on the other are assigned to the comparison group. It is considered a particularly useful design when researching programs that are given on the basis of need or merit and is considered a relatively robust design. In fact, inferences drawn from a well-implemented RD design are comparable in internal validity to conclusions from randomized experiments (Trochim, 2001). One of its strengths lies in the fact that it minimizes regression to the mean as a threat to internal validity by deliberately creating asymmetric samples that gives rise to expectation of regression towards the mean in both groups (Braden & Bryant, 1990). In effect, the RD design is a special from of regression analysis involving mutually exclusive groups. In sum, the RD design is a strong competitor to randomized designs when causal hypotheses are being investigated (Braden & Bryant, 1990; Marsh, 1998; Trochim, 2001).

In our RD design, participants were assigned to the TiPS or untreated comparison groups solely on the basis of a cutoff score on a pretest. Forty one students completed the TiPS curriculum

while the 161 students in the comparison group were left untreated. During this time, every effort was made to ensure that there are no intervening relevant instructional experiences for any of the participants, inside or outside of the study. At the conclusion of the study, we used design-appropriate statistical analysis to examine performance change across time—test performance before and after instruction—across both groups. Our analysis focused on determining (a) if there were learning gains associated with TiPS, (b) the typical duration of the TiPS instruction and whether the length of instruction varied in a systematic fashion according to a learner's mathematical ability prior to TiPS instruction, and (c) the nature of the knowledge that students acquired from TiPS.

Results

During Phase 1, TiPS reliably responded to a student who had achieved mastery at a certain level by encouraging the student forward to a new lesson by graying out items. Our analysis also suggested that TiPS was relatively tolerant to solutions that were conceptually accurate but deviated in some superficial way (e.g., different mapping, different tool selection) from the "expert" solution presented by TiPS. On the negative side, however, we discovered that the rate and level of hint use does not appear to inform the Bayesian Student Model. During Phase 2, we documented that the average score associated with the posttest for participants in the TiPS group was significantly higher than those of their peers in the untreated control group, indicating that there is a positive, practical effect associated with TiPS instruction. Essentially, we found that learners who spent an average of four hours on the TiPS system were typically rewarded with a 15% improvement—a 1½ letter grade improvement, by conventional standards—in their problem-solving performance. We also found that learners exposed to TiPS reported feeling that the instructional material, including the examples and problems, helped them to understand how to approach and solve word problems and that, overall, the instructional environment was well designed.

Significance

TiPS is intended for use by students, especially Navy personnel, who possess minimum competency with most basic arithmetic operations and thus are ready for applications-oriented instruction that focuses on helping students acquire and use basic mathematics knowledge as well as change and improve their problem-solving abilities in general. Students who have attained minimum competency in some, but not all, basic arithmetic operations, or those whose arithmetic skills are "rusty," also should be able to work successfully on TiPS because of its planned diagnostic and remediation features. Because the system's problem and domain information banks can be varied to meet a broad range of training needs, it is appropriate for use by the Navy in a variety of shipboard training programs, as well as in Jobs- and A-Schools. The system is also appropriate for use in middle schools, high schools, adult literacy programs, and workplace training programs. Results from several controlled field trials of the TiPS system suggest that adult remedial students can learn from the system and enjoy working with it. The results of this project replicate and extend these initial findings. Hopefully, one potential outcome of the project is the permanent installation, at several civilian and military sites, of a system for problem-based training in mathematics that reflects much current thinking in cognitive theory and research.

BACKGROUND

On TiPS, students receive instruction within the context of problem solving scenarios, designed to gradually build their skills and abilities that should enable them to reason about complex real-world problems. The instructional objectives of TiPS instruction include fostering everyday mathematics and problem-solving skills. In particular, these objectives include fostering the development of: (a) arithmetic schemas—conceptual structures shown by cognitive research to underlie human understanding of mathematics; (b) self-monitoring ability—the tendency and ability to be aware of one's own level of understanding and to check one's problem-solving performance to avoid careless errors; (c) supporting beliefs—the maintenance and use of beliefs associated with good problem solving; (d) selective encoding ability—the ability to identify important information and exclude extraneous information from the problem statement or situation; (e) strategic search ability—the ability to recognize the need for carrying out searches for problem information available through common indexed sources; and (f) strategic planning ability—the ability to select and organize schemas into solution steps that achieve an overall conceptualization of a solution to complex problems.

Despite the potential of TiPS to address a wide range of instructional objectives, several controlled field studies designed to test the pedagogical efficacy of the TiPS learning environment have generated results that are suggestive at best. One field study, which entailed high school students working with TiPS, identified that one important issue related to its effectiveness is the appropriate level of ability for students given TiPS instruction. Students that started with high math ability prior to the TiPS instruction—as evidenced by a high pretest score—did not appear to benefit from the system. On the other hand, there was evidence to suggest that students that struggled with the pretest and indicated that they were either in the middle or bottom third of the math classes they took in high school improved their performance on the posttest instruments. This represents a potential indicator of learning from the TiPS instruction. It is possible that these students derived greater benefit from the TiPS instruction, albeit this could also have been an example of regression towards the mean as well. Another field study conducted in conjunction with the Madison Area Technical College (MATC) adult literacy program did not generate significant gain scores (pretest to posttest differences). However, the results of a post-instruction questionnaire suggested that the participants enjoyed working with the system. In particular, the participants indicated that they liked the system's worked examples, thought the system helped them solve the posttest problems, and would recommend the system to a friend.

As noted in the Final Report for ONR N00014-93-1-0310 entitled "Development and Assessment of Tutorials in Problem Solving (TiPS): A Remedial Mathematics Tutor" by Sharon J. Derry and the TiPS Research Group, there were several drawbacks associated with the two field studies that limit the potency of their findings. Against this background, this project was designed to replicate and extend these initial findings.

TECHNICAL APPROACH

As previously mentioned, this project was designed to accomplish several objectives: (1) document the level of instructional adaptation provided by the TiPS system, (2) provide a basic evaluation of the overall TiPS system (3) determine the amount of instructional time involved with the use of TiPS system, and (4) determine the affective nature of the TiPS learning experience. To achieve these objectives, this project involved two phases: Phase 1 focused on addressing the first objective by relying on an analytic investigation of how the TiPS system behaves in response to various types of simulated performance; and Phase 2 addressed the remaining three objectives in the context of a regression-discontinuity (RD) design, a powerful pretest-posttest program-comparison group design that minimizes regression to the mean as a threat to internal validity. Each of these phases is described in detail below.

Phase 1: Analytic Investigation of the Tutor's Behavior

TiPS was designed to tailor itself to the individual learner based on its integrated Bayesian student model, which represents a theoretical model of problem solving knowledge. For instance, it was designed to adjust hints, evaluations scores, and other feedback provided during local problem solving; it adjusts the mastery level communicated to the student; and it may eliminate (indicated by grayed buttons) or add (indicated by active buttons) recommended practice problems. For example, a student who has achieved mastery at a certain level may be moved forward to a new lesson, or a remedial sequence may be suggested for a student that is having trouble.

During Phase 1, we conducted iterative user trials with systematic observations of system performance to help us document the level of instructional adaptation provided by the TiPS system, including verifying network functioning, that is, whether the Bayesian net registers beliefs about student problem solving that are consistent with the intuitive, common sense judgments of human teachers. This process enabled us to examine whether TiPS is doing any significant adaptation of instruction, including how much instructional variation it is actually producing. For instance, we predicted that the student model—if working properly—should catch people who do not need the additive portion of the TiPS instruction and (given adaptive instruction) move them through the curriculum expeditiously. Iterative user trials allowed us also to investigate the extent to which the key instructional decisions—such as which lessons to view—were decided by the software or are, conversely, under the learner's control. This is particularly important given that the nature and amount of learner control programmed into TiPS has not been sufficiently documented.

Sample and Design

To accomplish our experimental goals, we conducted iterative laboratory-based user trials with several graduate students. Essentially, the design consisted of a series of case studies where the students were systematically observed interacting with the TiPS system while adopting the behavior of "users" with a wide-range of ability levels. We acknowledge that the most obvious limitation of relying on these uncontrolled evaluations is the fact that the role of TiPS on

performance cannot be isolated unambiguously. However, this was the focus of Phase 2 of our project.

Computer-Based Learning Environment

The TiPS problem-solving interface was designed to help promote the learners' ability to model and reason about story problems (Derry, Wortham, Webb, & Jiang, 1996; Derry, Tookey, Smith, Potts, Wortham, & Michailidi, 1994; http://www.wcer/wisc/edu/tips/) and was built to adapt and extend Marshall's (1995) Story Problem Solver (SPS) interface. The SPS interface is designed to provide users with a small set of conceptually distinct diagrams for displaying and solving arithmetic problems. This approach represents a natural extension of classic expert/novice studies, which characterized good problem solvers as those possessing many conceptually rich knowledge structures, or "schemas," related to their domain of expertise. Marshall viewed schemas as abstract structures (instantiated either mentally or externally) that: (1) represent fundamental relational concepts within a domain; (2) suggest the existence of different problem classes; (3) suggest procedures associated with problem types; and (4) serve as conceptual building blocks for representing complex problems. The goal of SPS was to help students construct expert math knowledge by having them solve and analyze math story problems employing schematic diagrams representing basic semantic concepts.

The TiPS graphical user interface (GUI) supplies five schematic diagrams designed to serve as conceptual support for problem solving. Similar to SPS, each of the five TiPS core diagrams represents a different basic mathematics schema. The design for the diagrams was based on empirical evidence showing performance differences in problems solvable with the same arithmetic operations, but that have different semantics. The group relation, the comparison type, and change relation form a set of three additive classes found in arithmetic story problems. In addition, in both TiPS and SPS there are multiplicative situation classes as well, represented by restate (linear function) and vary (proportions) diagrams.

TiPS has a series of lessons associated with the goals of learning the five problem types and how to analyze and solve problems with their associated TiPS schemas. The system consists of two sets of lessons: (1) basic schema tool lessons and (2) advanced lessons. In the basic schema tool lessons, there is one instructional unit each for the Change, Group, Compare, Vary and Restate tools. With the exception of the Vary unit, each instructional unit consists of one lesson. There are four lessons associated with the Vary unit (i.e., proportion, rate, percent, and "slice of life problems"). Thus, there are a total of eight lessons that make up the basic schema tool lessons. The advance lessons consist of two lessons, one that involves a set of mixed one-step problems representing the five problem types and another that involves multi-step problems. In the mixed one-step advanced lesson, the practice set requires students to discriminate among all five problem types for tool schema selection. The multi-step advanced lesson focuses on advanced problem solving, including problem solving practice with complex, multi-step problems[1].

Within these lessons, students study dynamic worked examples that illustrate expert problem solving on TiPS, and they complete practice problems that are similar to the worked examples.

[1] Note: TiPS provides minimal feedback with no on line cognitive diagnosis for this lesson since the system does not incorporate an AI algorithm for multi-step cognitive diagnosis.

The worked examples illustrate desired problem solving performance with didactic audio explanation from a tutor. The worked examples incorporate the schematic diagrams used in the problem-solving interface. Thus, they are designed to illustrate a schema-based approach to problem solving. The lessons also include a set of 6-9 regular practice problems plus up to 4 optional skill building practice problems.

Students receive hints and evaluative feedback designed to help them learn the problem solving skills associated with each lesson. These hints are provided by a local evaluation component of TiPS. The TiPS system also has a global evaluation component. These two components of evaluation are independent yet interrelated. The local evaluation component is concerned with accurate diagnosis and feedback on a particular problem whereas the global evaluation component is concerned with building a picture of overall problem solving competency and behaviors of the student over time. The local evaluation component is provided by the cognitive diagnoser described previously. The heart of the global evaluation component is represented by a Bayesian network. Both the local and the global evaluation components work together. Based on its integrated Bayesian student model, TiPS is capable of adapting the system to tailor its actions in several ways. It can adjust hints, evaluation scores, and other feedback provided during local problem solving; it is capable of adjusting the mastery level communicated to the student; and it may eliminate (indicated by grayed buttons) or add (indicated by active buttons) recommended practice problems.

Procedure

As previously mentioned, the graduate students interacted with the TiPS system while simulating the behavior of users with a wide-range of ability levels and the progress through the system was systematically observed. We started the process simulating students on the extremes of the ability continuum and observing how TiPS adapts its instruction accordingly. For instance, we examined what happens if a "student" performs perfectly—does that student get asked to solve every problem, get rapidly advanced to the "advanced" lessons or not. We also examined what happens if a simulated student makes lots of errors, systematic (on a certain type of problem) or not. We continued this process by introducing simulated users with less extreme ability levels until we documented under what conditions the Bayesian inference network was able to adjust rapidly to the characteristics of whatever student it encounters.

Results

The tutorial program was completed under six different circumstances with a focus on different parameters for the various runs. The first two runs were designed simply to determine how the program reacts to students who work all problems correctly versus students who make continuous errors. The third run focused on discrepancies in scoring, primarily due to mislabeling, entering data in the wrong use of hints. The final two runs were conducted by students and recorded using pcAnywhere software (see Appendix A for anecdotal comments regarding TiPS from graduate students). A careful analysis of these two recordings was then conducted to determine whether any additional problems evident with the system.

All Problems Worked Correctly

Across four different trials, the program began with **80** problems to work (i.e. not "grayed out") and *consistently* reduced this requirement throughout the curriculum as the user correctly worked the problems until only a total of **58** problems were required to be solved by the user.

Lesson Type	*Problems to Work*	*Total*
Change	6 regular, 4 practice	10
Compare	2 regular, 4 practice	6
Group	2 regular, 4 practice	6
Function	4 regular, 4 practice	8
Vary 1	2 regular, 3 practice	5
Vary 2	2 regular, 4 practice	6
Vary 3	2 regular, 3 practice	5
Vary 4	0 regular, 0 practice	0
Advanced 1	6 regular, 0 practice	6
Advanced 2	6 regular, 0 practice	6
	Total:	**58**

All Problems Worked Incorrectly

Across four different trials, the program began by requiring the user to solve **80** problems (i.e. not "grayed out") and *consistently* maintained this requirement throughout the curriculum until all **80** problems were solved by the user.

Lesson Type	*Problems to Work*	*Total*
Change	6 regular, 4 practice	10
Compare	4 regular, 4 practice	8
Group	4 regular, 4 practice	8
Function	6 regular, 4 practice	10

Vary 1	6 regular, 3 practice	9
Vary 2	4 regular, 4 practice	8
Vary 3	6 regular, 3 practice	9
Vary 4	6 regular, 0 practice	6
Advanced 1	6 regular, 0 practice	6
Advanced 2	6 regular, 0 practice	6
	Total:	**80**

All Problem Worked Correctly but Answers Varied from TiPS Solution

On a single trial, the program began with **80** problems to work (i.e. not "grayed out") and finished with only **61** problems that required working.

Lesson Type	*Problems to Work*	*Total*
Change	6 regular, 4 practice	10
Compare	3 regular, 4 practice	7
Group	2 regular, 4 practice	6
Function	4 regular, 4 practice	8
Vary 1	3 regular, 3 practice	6
Vary 2	2 regular, 4 practice	6
Vary 3	2 regular, 4 practice	6
Vary 4	0 regular, 0 practice	0
Advanced 1	6 regular, 0 practice	6
Advanced 2	6 regular, 0 practice	6
	Total:	**61**

Example of "Improper" Tool Selection

Item in Curriculum:

> "The Lemmon Sisters' Cookie Company was now selling over 1,300 cookies a day. Last week they made $1,075. This week, their income increased by $205, a new company record! What was their income this week?" **[Change Lesson, Problem C2]**

Student Solution and System's Response:

> The Compare Tool was used to solve the problem, comparing $1,075 for last week with an increase of $205 for this week, or $1,075 + $205 = $1,280 for this week. TiPS, however, expected the Change Tool to be used, using $1,075 as the starting amount, $205 as the ending amount (?) and the change amount as $1,280. Because an "improper tool" was used to solve the problem, TiPS gave a score of 1/10. This score seemed quite harsh considering the correct answer was obtained. In fact, it could be argued that the Compare Tool was as logical in this problem as the Change Tool. Furthermore, it would seem that the TiPS solution should have had $205 entered as the change amount and $1,280 as the ending amount to make the Change Tool selection make sense.

Example of Improper Mapping

Item in Curriculum:

> "When purchasing some notebooks you hand the cashier a $10 bill and receive $3.86 in change. What was the total price of the notebooks?" **[Change Lesson, Practice Problem 3]**

Student Solution and System's Response:

> The Change Tool (correct choice) was used for this problem. However, the entries were $10 for the start amount, $3.86 for the change amount resulting in $6.14 for the ending amount (price of notebooks: $10 - $3.86). TiPS deducted one point for using the wrong boxes. The "correct" procedure was to enter $10 as the start amount, $3.86 as the ending amount and the change amount to be $6.14 (the difference between the two amounts is the price of the notebooks.)

Example of Failing to Label Sets Properly

Item in Curriculum:

> "The speed of sound in air at a temperature of 20 degrees Celcius is approximately 350 m/s meters per second), while the speed of sound in water is 1500 m/s. What is the difference between the speed of sound in water and the speed of sound in air in m/s?" **[Compare Lesson, Practice Problem 3]**

Student Solution and System's Response:

The Compare Tool (correct tool) was used. However, the labels used were 350 m/s for the "a amount", 1500 m/s for the "b amount" and 1150 m/s as the "compare amount." TiPS expected the labels to read 350 m/s in air as the "a amount", 1500 m/s in water as the "b amount" and 1150 m/s difference as the "compare amount." As a result, one point was deducted for mislabeling the problem.

Examples of Proportion Problem "Glitches"

(Example 1) Item in Curriculum:

"Duke: Rather than taking our chances at the BlackJack table, we could just lay low and save up the money we earn from our vaudeville act. After all, back in San Francisco we earned about $12 for every 5 shows. How many shows would it take for Duke and Chester to earn the $750?" **[Vary Lesson 1, Problem A2]**

(Example 1) Student Solution and System's Response:

The proportion was set up as follows: "If 5 shows = $12 earned, Then how many shows = $750 to earn?" The answer was obtained by multiplying 750 times 5, then dividing by 12 to get 312.5 shows. The TiPS solution showed the exact same procedure; however, 3 points were deducted on the score. When the problem was reworked, putting the dollars earned on the left side of the proportion and the number of shows on the right side, still giving the same formula with the same answer, TiPS scored the reworked problem 10/10.

(Example 2) Item in Curriculum:

Vacation days are earned at the rate of 1.5 days for every 2 full months of employment. How many months of service are needed to take a total of 13 days of vacation? **[Vary Lesson 1, Practice Problem 3]**

(Example 2) Student Solution and System's Response:

The proportion was set up to read "If 2 months employment = 1.5 days vacation, then how many months employment = 13 days vacation?" The answer obtained by multiplying 13 x 2 and dividing by 1.5, giving 17.33 days, was correct but was scored as 6/10. TiPS, instead, looked for the proportion to read: "If 1.5 days = 2 months employment, then 13 days = how many months employment." This revised set-up would have no effect on the solution procedure or answer. Therefore, the deduction of points is inappropriate.

(Example 3) Item in Curriculum:

Sodium hypochlorite (household bleach) is said to be a good biological decontaminate. When used for decontamination of clothing it should be diluted (2/3 cup of bleach per 1

gallon { gal(s) } of water). If 4 gallons of water were required for a mixture to decontaminate a large bundle of clothing, how many cups of bleach should be added to the water?" **[Vary Lesson 2, Practice Problem 1]**

(Example 3) Student Solution and System's Response:

In this problem, the proportion was set up to read "If 1 gallon of water = 2/3 cup of bleach, then 4 gallons of water = how many cups of bleach?" The solution was found by multiplying 2/3 times 4, giving 2.667 cups as the answer. TiPS set it up the same but worked it by multiplying 2/3 times 4, then dividing by 1. This still gave the same answer, but TiPS deducted 3 points for failing to divide by 1.

(Example 4) Item in Curriculum:

"Bill's favorite salad dressing uses 4 tablespoons (tbl) of sour cream per 1 teaspoon (tsp) of vinegar. Bill wants to make a mixture that tastes the same but that uses 3 tbl of sour cream. How much vinegar will Bill need to use?" **[Vary Lesson 2, Practice Problem 2]**

(Example 4) Student Solution and System's Response:

This problem was set up as follows: "If 1 tsp of vinegar = 4 tbl of sour cream, then ? tsp vinegar = 3 tbl of sour cream?" The solution was obtained by dividing 3 by 4 and getting .75 tsp vinegar. The TiPS solution set the problem up the same, but obtained the answer by multiplying 3 time 1 before dividing by 4. This did not change the answer, but the penalty for not multiplying by 1 was a 3 point reduction in the score.

All Problems Worked Correctly but Hints Used Excessively

Across four different trials, the program began with **80** problems to work (i.e. not "grayed out") and finished with only **54** problems that required working (due mainly to the fact that Vary Lesson 4 as well as both Advanced lessons were completely grayed out.)

Lesson Type	*Problems to Work*	*Total*
Change	6 regular, 4 practice	10
Compare	4 regular, 4 practice	8
Group	2 regular, 4 practice	6
Function	5 regular, 4 practice	9
Vary 1	5 regular, 3 practice	8

Vary 2	4 regular, 4 practice	8
Vary 3	2 regular, 3 practice	5
Vary 4	0 regular, 0 practice	0
Advanced 1	0 regular, 0 practice	0
Advanced 2	0 regular, 0 practice	0
	Total:	**54**

"Average-to-High" Ability Student

On a single trial, the program began with **80** problems to work (i.e. not "grayed out") and finished with **63** problems that required working.

Lesson Type	*Problems to Work*	*Total*
Change	6 regular, 4 practice	10
Compare	4 regular, 4 practice	8
Group	2 regular, 4 practice	6
Function	6 regular, 4 practice	10
Vary 1	3 regular, 3 practice	6
Vary 2	2 regular, 4 practice	6
Vary 3	2 regular, 3 practice	5
Vary 4	0 regular, 0 practice	0
Advanced 1	6 regular, 0 practice	6
Advanced 2	6 regular, 0 practice	6
	Total:	**63**

Summary of Performance in Additive Lessons

Change Lesson

Problem A2: Correct answer was obtained but set-up was incorrect. **Score 8/10.**

Note: Student was only required to work 2 out of 6 regular problems and 1 out of 3 practice problems in this lesson. Of those worked, one (practice problem) was worked incorrectly.

Compare Lesson

Practice Problem 1: Correct answer, but mislabeled parts of the problem. **Score 8/10**.

Note: Missed 1 of the remaining 7 regular and practice problems.

Group Lesson

Practice Problem 3: Correct answer, but mislabeled one entry in the problem. **Score 6/10**.

Note: The other 5 regular and practice problems were worked correctly.

Summary of Performance in Multiplicative Lessons

Function Lesson

Problem B1: Correct answer, but incorrect set-up. **Score 8/10**.

Note: Missed 2 of the remaining 9 regular and practice problems.

Vary Lesson 1

Problem A1: Answer far off due to failure to divide by 20. **Score 9/10**

Practice Problem 3: (worked before Practice Problem 2): Set up and worked correctly and received correct answer. **Score 6/10**.

Note: TiPS appeared to drop a digit upon transferring a number (i.e. 13 became a 3). However, it still calculated it using the correct number. Then in Practice Problem 2, TiPS again appeared to drop a digit (i.e. 18.5 became 8.5), but this time the student received a correct score of 10/10.

Vary Lesson 2

Problem A1: Set up incorrectly, answer far off, **score 7/10.**

Problem B1: Worked correctly, correct answer, **score 9/10.** ***(Anomaly in System)***

Practice Problem 1: Student entered numbers to calculate as 4 times 1 divided by two thirds (which would give 6). However TiPS calculated it as 4 times 1 divided by 2, divided by 3 (which gave .67). Consequently, there appears to be a problem concerning Order of Operations. The problem was then reworked, setting it up as 4 divided by 1 times 2 divided by 3 and the correct answer was obtained (i.e. 2.667 cups of bleach). However, the score for the reworked problem was only **7/10**. The TiPS solution said to multiply by two thirds first and then divide by one, but this would make no difference in the answer.

Practice Problem 2: Student worked the problem exactly like the solution, getting the correct answer. Yet the score was **7/10**.

Practice Problem 3: Set up incorrectly and worked incorrectly, but correct answer still obtained (with wrong units). **Score 8/10.**

"Average-to-Low" Ability Student

On a single trial, the program began with **80** problems to work (i.e. not "grayed out") and finished with **73** problems that required working.

Lesson Type	*Problems to Work*	*Total*
Change	6 regular, 4 practice	10
Compare	4 regular, 4 practice	8
Group	2 regular, 4 practice	6
Function	6 regular, 4 practice	10
Vary 1	6 regular, 3 practice	9
Vary 2	3 regular, 4 practice	7
Vary 3	5 regular, 3 practice	8
Vary 4	3 regular, 0 practice	3

Advanced 1	6 regular, 0 practice	6
Advanced 2	6 regular, 0 practice	6
	Total:	**73**

Summary of Performance in Additive Lessons

Change Lesson

Problem B1: Set up correctly but failed to make the calculation. **Score 8/10**.

Problem C1: Set up correctly but failed to make the calculation. **Score 8/10**.

Practice Problem 3: Start and end numbers were reversed, giving the wrong sign for the answer. **Score 4/10**.

Practice Problem 4: Wrong numbers used in calculation. **Score 6/10**.

Compare Lesson

Problem A1: Correct numerical answer, wrong set-up/labeling. **Score 8/10**.

Problem B2: Set up correctly but failed to make the calculation. **Score 8/10**.

Practice Problem 1: Used the wrong tool (Change tool instead of Compare tool). Otherwise the problems was set up and worked correctly with right answer and correct units. **Score 8/10**.

Practice Problem 2: Used the Change tool to work, again, instead of the Compare tool. Set the problem up logically, all correct units, worked it correctly and obtained the correct answer. **Score 1/10**. ***(Anomaly in System)***

Group Lesson

Note: Total number of grayed-out problems was increased from 6 to 8, leaving six problems to work rather than eight. All problems in this section were worked correctly with all scores **10/10**.

Summary of Performance in Multiplicative Lessons

Function Lesson

Problem A1: Set up correctly but failed to make the calculation. **Score 9/10.**

Problem A2: Set up problem correctly, except for the inclusion of a multiplication sign, but failed to make the calculation. **Score 8/10.**

Problem A3: Set up correctly, but failed to make the calculation. **Score 8/10.**

Problem B1: Set up correctly but failed to make the calculation. **Score 7/10.**

Problem B2: Set up correctly, then made the wrong calculation (subtracted instead of divided) and got an answer far from the correct answer. **Score 8/10.**

Problem B3: Set up incorrectly which would have led to an answer of 9,232,600 (when correct answer was 12.170), then failed to solve. **Score 7/10.**

Practice Problem 1: Set up correctly but failed to calculate. **Score 9/10.**

Practice Problem 2: Set up wrong, worked wrong, wrong answer. **Score 6/10.**

Practice Problem 3: Set up wrong, worked wrong, wrong answer. **Score 7/10.**

Practice Problem 4: Set up correctly, worked wrong (multiplied instead of divided) and obtained answer far off. **Score 8/10.**

Vary Lesson 1

Problem A2: Selected correct tool, set up problem correctly, solved correctly and obtained the correct answer. **Score 7/10.** ***(Anomaly in System)***

Problem A3: Failed to set up or work. Only checked solution. **Score 2/10.**

Problem B1: Set up correctly, but failed to calculate. **Score 7/10.**

Problem B2: Set up correctly, but failed to calculate. **Score 8/10.**

Vary Lesson 2:

Note: Grayed out problems increased from 6 to 7, leaving 7 problems to work instead of 8. Also, the student did not complete the 4 practice problems.

Problem B1: Set up correctly, worked correctly and correct answer obtained. **Score 9/10.** ***(Anomaly in System)***

Vary Lesson 3

Note: Grayed out problems increased from 5 to 6, leaving eight problems to work instead of nine.

Problem A2: Set up problem the same as the solution set-up (but, not logically, in my opinion, since it begins with "If 5% = 100%"). Failed to calculate the answer. **Score 5/10**.

Problem B1: Set up problem wrong (plus illogically), then worked properly, but wrong answer. **Score 8/10**.

Problem B2: Set up problem the same as the solution, but failed to include the % sign in the answer. **Score 9/10**. (Note: Problem B3 was grayed out here.)

Vary Lesson 4

Note: Following Problem A3 (score 10/10), all three problems in the B column were grayed out, reducing the total of problems to work from 6 to 3.

Problem A1: Set up problem incorrectly, wrong answer. **Score 7/10**. ***(Anomaly in System)***

Problem A2: Set up problem incorrectly, wrong answer. **Score 8/10**. ***(Anomaly in System)***

Advanced Lessons 1

Problem A3: Set up correctly but failed to calculate. **Score 8/10.**

Problem B3: Set up correctly but failed to calculate. **Score 8/10**.

Advanced Lessons 2

Note: Student attempted two of the six available problems and part of a third problem, but all were incorrect.

Discussion and Conclusions

The results of the first two runs—All Problems Worked Correctly and All Problems Worked Incorrectly—provided clear evidence that the Bayesian Student Model does monitor student performance and respond appropriately by tailoring instruction, as indicated by the significant reduction in the number of required problems for students who work the problems correctly. As previously indicated, TiPS is designed to automatically "gray out" problems at a given level once it has determined that a student has achieved mastery at that level, thereby effectively reducing the required number of problems to work. A comparison of the total number of "Problems to Work" for these two cases shows a decrease of 22 problems (from 80 to 58) for the student

working all problems correctly versus no decrease for the student working all problems incorrectly. Therefore, the design feature for graying out problems based upon topic mastery seems to work properly.

The third run—All Problem Worked Correctly but Answers Varied from TiPS Solution—demonstrated how scoring was affected due to a variety of factors, including "improper" tool selection, improper mapping, and failing to label set properly. Basically, we were seeking to identify the impact of submitting answers that appeared superficially (but not conceptually) different than what TiPS expected, such as using a different tool than the one TiPS considered to be the best tool or mapping the sets differently from what TiPS expected. All problems in this run were still set up logically and worked correctly in the sense that the correct answers were obtained. As before, due to the accuracy of the solutions, TiPS decreased the number of problems to work by graying out 19 problems (from 80 to 61). This provided additional evidence that TiPS is capable of recognizing a student's level of mastery and making instructional adaptations accordingly. However, on several occasions, we found that the scoring penalty was sometimes quite harsh (up to 9 out of 10 points deducted) for using a tool other than what TiPS expected, even though that tool might still be considered a logical tool (e.g., a group tool instead of a compare tool) for the problem. On the other hand, the penalty for mapping the set into the "wrong" part of the diagram or for failing to label sets properly in the problem setup was usually minor (e.g., one point). Overall, the system does not appear to penalize the user too dramatically—in this instance, only requiring the user to solve three additional problems (i.e., 61 instead of 58)—for varying from the "expert" solution in TiPS.

It is, however, noteworthy that some of the Vary Lesson problems in the third run had various "glitches" that appear to make the TiPS scoring particularly unpredictable. For example, proportions set up correctly, but different than the TiPS solution model, were often docked 3 points, even though still worked logically to obtain the correct answer. Other problems had 3 points deducted for not multiplying the answer by 1 or dividing the answer by 1 (which would have no effect on the answer).

The fourth run—All Problems Worked Correctly but Hints Used Excessively—examined how TiPS reacted in terms of scoring and graying out problems when all three levels of hints were repeatedly used to obtain the correct answer (compared to the situation in the first run where all of the answers were correct and no hints were used). Basically, the hint button in TiPS allows a user to solve each problem correctly using a "brainless" approach if he or she wanted to by simply clicking on the button repeatedly in order to obtain "hints" that illustrate how to correctly solve each of the individual problems in the curriculum. Distressingly, when compared to the first run in which no hints were used, both had identical mastery scores (perfect). Moreover, the "graying out" feature of TiPS differed across the first and fourth runs. For the first eight lessons (Change lesson through Vary 4 lesson), TiPS grayed out 8 more problems on the run that used no hints. Then on the ninth and tenth lessons (Advanced 1 and 2), TiPS grayed out both lessons completely for the run using hints while leaving 12 problems (6 in each advanced lesson) to work for the run that didn't use the hints. The final tally, therefore, is deceiving in that it indicates the "brainless approach" use of excessive hints gives a perfect mastery score and requires the least number of overall problems to work. In a worst case scenario, a student could

complete the entire tutorial and achieve a perfect score without having read any of the problems and without having learned anything about their solutions.

The last two runs—"Average-to-High" Ability Student and "Average-to-Low" Ability Student—were actual test runs by students. The first student had a stronger background in math than the second and the comparison of the numbers of problems grayed out captures the difference in ability. As with the third run, several instances were discovered where the student worked every aspect of the problem correctly and still had points deducted. Also, we noticed that TiPS appears to have a problem following proper order of operations in problems that involve fractions (1st Student, Vary Lesson 2, Practice Problem 1).

In sum, on the positive side, certain aspects Bayesian Student Model worked properly. TiPS did consistently adjust the mastery level communicated to the student and eliminated (as indicated by grayed buttons) and added (as indicated by active buttons) recommended problems and practice problems. For instance, TiPS reliably responded to a student who had achieved mastery at a certain level by encouraging the student forward to a new lesson by graying out items. This analysis also suggested that TiPS was relatively tolerant to solutions that were conceptually accurate but deviated in some superficial way (e.g., different mapping, different tool selection) from the "expert" solution presented by TiPS. On the negative side, the most glaring problem is the fact that the rate and level of hint use does not appear to inform the Bayesian Student Model. Instead, a student could use hints to complete the entire tutorial, while along the way be required to solve the least number of overall problems and achieve a perfect score—all without having to cogitate on a single problem.

Phase 2: Laboratory-Based Evaluation of TiPS

This phase is designed to accomplish three of the proposed project's four main objectives, including (1) provide a basic evaluation of the overall TiPS system, (2) determine the amount of instructional time involved with the use of TiPS system, and (3) determine the affective nature of the TiPS learning experience. To accomplish this task, we relied on a RD design, a powerful methodological alternative to quasi-experimental or randomized experiments when conducting evaluations of education programs. According to Braden and Bryant (1990), a RD design is "among the strongest models for testing program efficiency when selection into a program is based on a continuous criterion and random assignment is not possible [and that] other alternatives (e.g., contrasting pretest to posttest 'gain' between selected and excluded group) are susceptible to regression to the mean, attrition, and changes in the interval between selection and outcome testing" (p. 234). We used the RD design to empirically examine whether differences exist between the treatment group (TiPS users) and an untreated comparison group (non-TiPS users). The actual implementation of the RD design is described below.

Sample and Design

The laboratory-based experiment was conducted with adult remedial volunteers drawn from Mississippi State University that appeared to match the achievement and aptitude profile of the Naval recruits who would most likely use the TiPS program (e.g., naval recruits in the JOBS program at the Great Lakes Naval Training Center in Chicago, IL). According to the admissions

office at Mississippi State University (MSU)—a land-grant institution with a 71% acceptance rate and where the average composite ACT score falls between 19-27, many of the young adults entering the institution as undergraduates demonstrate inadequate readiness in English, reading, or mathematics. As a result, these students are required to participate in remedial programs offered on campus. Thus, at least a portion of the undergraduate population at MSU appeared comparable to the Navy recruits considered to be in need of remedial instruction

The "basic" RD design is a pretest-posttest, two-group design. The term "pretest-posttest" implies that the same basic measure—or in our case, alternative forms of the same measure—is administered before and after a program or treatment. The key feature of the RD design is assignment to the program is based on a cutoff value on the pretest, where the cutoff rule is essentially: (1) all persons on one side of the cutoff are assigned to one group, (2) all persons on the other side of the cutoff are assigned to the other, and (3) there needs to be a continuous quantitative pre-program measure (i.e., pretest). In this case, the selection of the cutoff was made on the basis of a pilot study.

Since the general rule of thumb for the RD design is 30 or more subjects in the program group (i.e., TiPS condition), with at least twice as many in the excluded group, we decided to test a large number of students to ensure that a sufficient number would fall below our designated cutoff. As a precaution, we made arrangements to test approximately 300 students. To ensure adequate participation in the study, the initial group of volunteers was offered $5.00 for taking a pretest and an additional $20.00 if they were willing to return and take the posttest at a later date. Of the students eligible for TiPS (i.e., students that score below our criterion on the pretest were eligible to work on the TiPS system), those that elected to participate in the treatment portion of the study were paid a pre-determined amount for each training module they completed Specifically, the TiPS system consists of ten lessons. Participants were paid $8 for each of the first four stages they completed, $10 for each of the next four stages, and $12 for each of the final two stages.

Two hundred and eighty three students were administered the pretest. Of the 283, 124 students—or 44%—scored below the cut-off score (earning a score of 40% or less) on the pretest. Of the 124 students, we randomly selected half of them to approach and notify that they were eligible for the treatment (i.e., TiPS). Of the 62 we approached, 43 (35%) expressed a willingness to participate in the treatment portion of the study; the other students elected to drop altogether from the study. Two of these students that initially expressed interest, however, did not complete the TiPS instruction and were excluded from the final analysis. Thus, a total of 41 students completed the entire TiPS curriculum. Of the remaining 62 students that scored below the cut-off score on the pretest but were not selected, 41 took the posttest and the other 21 dropped completely out of the study (they did not take the posttest). Of the 159 students that scored above the criterion, 116 took the posttest while the remaining students dropped completely out of the study (they did not take the posttest). In sum, there were three groups of students that participated in the entire study and were used in the final analysis:

- **41** students (14 males and 27 females) in the **TiPS treatment group** (below cut-off on pretest). The average ACT math score for these students was 18.10 (SD = 3.07). Of the 41 participants in this condition, 19 were Caucasian, 21 were African American, and 1 described himself as "other." Eleven of the 41 participants classified themselves as

lower-division undergraduates with the rest classifying themselves as upper-division undergraduates.

- **41** students (11 males and 30 females) that were the eligible for TiPS (below cut-off on pretest) but were not randomly selected to receive TiPS treatment (i.e., **eligible/untreated group**). The average ACT math score for these students was 18.36 (*SD* = 3.18).). Of the 41 participants in this condition, 19 were Caucasian, 20 were African American, and 2 described themselves as "other." Thirteen of the 41 participants classified themselves as lower-division undergraduates with the rest classifying themselves as upper-division undergraduates.
- **116** students (42 males and 73 females) in the comparison or **untreated control group** (above cut-off on pretest). The average ACT math score for these students was 22.80 (*SD* = 5.03). Of the 116 participants in this condition, 89 were Caucasian, 17 were African American, 1 was Hispanic, 2 were Asian American, and 7 described themselves as "other." Thirty-seven of the 116 participants classified themselves as lower-division undergraduates with the rest classifying themselves as upper-division undergraduates.

Computer-Based Learning Environment

The computer-based learning environment (TiPS) used during Phase 2 was the same as the one used during Phase 1.

Materials

The pencil-paper materials included a demographic questionnaire, a pretest, a posttest, and an affective questionnaire. The demographic questionnaire asked each learner to provide information (e.g., standardized test scores, ethnicity, gender).

To determine the effects of TiPS instruction on word problem performance, a pretest and posttest was administered to each student immediately prior to and following instruction. The pretest and posttest was adapted from a set originally developed by Derry and her students to evaluate the system. Based on prior research with TiPS (Wortham, 1996), adult remedial learners typically enter TiPS instruction already performing well on one-step change, group, and compare (additive) word problems, but not on one-step vary and function (multiplicative) word problems or on multi-step problems involving both multiplicative and additive schemas. To obtain an instrument that would allow one to measure the instructional impact and that could be completed in a reasonable time period (i.e., one hour), two eight-item tests were created each consisting of four one-step multiplicative word problems (involving the vary and function schemas) and six multi-step word problems involving both multiplicative and additive schemas. These tests were based on the tests used in two field studies described previously. Two equivalent forms (A and B) of a test were developed (see Appendix B), each consisting of four one-step, three two-step, and three three-step word problems. With respect to mathematics operations, underlying concepts, sizes of numbers, and basic grammatical features such as sentence structure, problems on the two forms were structurally isomorphic to one another. Statistical treatments were designed to assess whether the instruments perform similarly. One single-step and one two-step problem on each test were obtained from the National Assessment of Education Progress

(NAEP). Test administration was counterbalanced so that half the students received form A prior to instruction and form B following instruction.

An affective questionnaire was also created that asked each participant to judge the effectiveness of the instructional program. Specifically, the questionnaire consisted of a set of seven statements to which the participants responded on a 5-option Likert-type scale from "I disagree" to "I agree". For instance, the following statements may be used: (1) "I have learned to solve word problems based on this instruction"; (2) "Learning was fun"; (3) "I would prefer learning from TiPS when I have to study "mathematized" contents next time; (4) "I felt curious"; (5) "The examples and problems in TiPS helped me to understand word problems"; (6) "I was interested in learning about word problems"; and (7) "The instruction in TiPS was well designed". The questionnaire was designed to contain a balance of negatively and positively worded items.

Procedure

In order to identify at least thirty students that would fall below the cutoff—and thus be deemed eligible for TiPS instruction, a large pool of participants were recruited as volunteers from remedial mathematics classes at the Mississippi State University. As previously mentioned, the selection of the cutoff was made on the basis of a pilot study. All participants received cash payments for their participation based on their level of participation. During the Pretest Session, the volunteers were requested to provide human subjects consent and then asked to complete the pretest.

Based on the results of the Pretest Session, individuals that scored below the cut-off score were identified and asked to return for subsequent sessions. As previously mentioned, to help ensure that these individuals were sufficiently motivated, we provided them payments for work completed. Also, at this point, every effort was made to ensure that there are no intervening relevant instructional experiences, inside or outside of the study.

For the TiPS-based treatment sessions, experimenters followed an invariant data collection protocol that they were trained to employ. The steps in this protocol included: 1. Preliminary preparation (e.g., readying the computer, ensuring students properly log into TiPS); 2. Administration of TiPS instruction; and 3. Administration of posttest and affective questionnaire. Since the instruction was self-paced, the actual length of time necessary to complete the data collection protocol varied across participants. To ensure that the individual sessions during step 2 (i.e., administration of instruction) of the protocol do not get too long, the participants were not be permitted to work with the system for more than an hour a day.

The students in the comparison group—that is, the students with scores above the cutoff on the pretest—and the students in the eligible/no treatment group that participated in the Pretest Session were asked to return for the Posttest Session in which the posttest was administered on either an individual or group basis. This session were timed to coincide with the last session of the participants exposed to the TiPS system.

Scoring

The protocols generated on the pretest and on the posttest were coded for conceptual scores according to a set of guidelines (see Appendix C) for analyzing the written problem-solving protocols derived from research by Derry and her students (Atkinson & Derry, 2000; Derry, Weaver, Liou, Barker, & Salazar, 1991; Tookey & Derry, 1994). These guidelines were designed to help gauge where the participant fell along a problem-comprehension continuum. According to these guidelines, each item will be awarded a conceptual score, ranging from 0 to 3, depending upon the degree to which the participant's solution is conceptually accurate. Thus, the overall score for both the pretest and the posttest ranged from 0 to 30. The cutoff criterion was 40% or a score of 12 on the pretest.

One research assistant who was unaware or "blind" to the condition independently coded each protocol. To validate the scoring system, two raters independently scored a random sample of 20% of the problem-solving protocols and agreed on scoring 97% of the time. Discussion and common consent were used to resolve any disagreement between coders. Once the pretests and posttests were scored, gain scores were calculated designed to capture any pretest-to-posttest differences.

To create an average affective score, the participants' responses to all of the questionnaire items were coded on a scale of 1 to 5. The participants' responses will then be summed across all of the seven questions and divided by seven, thereby generating an average response on the affective measure, with values ranging from 1 to 5.

Analysis

The unadjusted pretest and posttest scores for the three conditions appear in Table 1. To determine if there are learning gains associated with TiPS (i.e., Objective 1 – provide a basic evaluation of the overall TiPS system), the analysis of problem-solving measures consisted of two complimentary alpha-controlled sets of analyses. First, an analysis of covariance (ANCOVA) was used for testing regression discontinuity effects (Braden & Bryant, 1990; Cook & Cambell, 1979). We entered pretest scores as the covariate, entered placement (TiPS group or comparison group) as the independent variable, and designated posttest scores as the dependent variable in the ANCOVA. With this approach, the difference or discontinuity at the cutting point between the regression surfaces in the two groups can be taken as evidence of a treatment effect. The interaction between pretest and placement was also entered to test whether there is or is not a difference in slope between the two groups (i.e., homogeneity of regression lines).

Second, since it is possible for the ANCOVA to be misspecified such that the shape of the regression surface is not properly modeled—for instance if there is a curvilinear relationship between the pretest and posttest, we attempted to exactly specify the true model. When we exactly specify the true model, we obtain unbiased and efficient estimates of the treatment effect. Our general strategy was to begin specifying a model that we are fairly certain was overspecified. Although the treatment effect estimate for this initial model was likely to be unbiased, it was also considered inefficient. Through successive analyses, we gradually removed higher-order terms until the model diagnostics indicate that the model fits poorly. Specifically,

the basic model specification analysis for RD designs involves five steps: (1) transform the pretest, (2) examine relationship between pretest and posttest visually, (3) specify higher-order terms and interactions, (4) estimate initial model, and (5) refine the model (Trochim, 2001).

We also focused our attention on the participants that interacted with the TiPS system in order to address the remaining two objectives (i.e., determining the amount of instructional time involved with the use of TiPS system and determine the affective nature of the TiPS learning experience.). For instance, we calculated the instructional time and the number of problems the students solved during instruction. In addition, we examined the posttest questionnaire for evidence that the participant was affected by the TiPS instruction. In addition,

Finally, we also examined the relationship between those students that were eligible—by scoring below the cutoff criterion—to participate in the treatment portion of the study but randomly selected to not participate in the treatment (eligible/non-participants) and the students that were eligible and did participate in the treatment. Essentially, the eligible/non-participants formed a second comparison group that could be used to examine the pattern of performance from pretest to posttest that one is likely to see among low performing students.

Results and Discussion

Comparing TiPS Group to Untreated Control Group

Figure 7 shows the bivariate distribution between the pretest and posttest for this experiment. Each dot on the figure represents an individual student's pretest and posttest. The vertical line that appears at the pretest score of 12 on the x-axis represents the cutoff criterion. The dashed lines that appear through the bivariate distributions on both sides of the cutoff score are the regression lines associated with the TiPS group (on the left of the figure) and the control group (on the right of the figure). On the basis of a visual inspection of Figure 7, one can perceive a "jump" or discontinuity in the regression lines at the cutoff point. Specifically, it appears that—on average—the points to the left of the cutoff (i.e., the TiPS treatment group) have been raised by approximately 4 points on the posttest.

Although one might conclude from a visual inspection of Figure 7 that TiPS on average raised posttest performance by 4 points on our scale, we wanted to confirm it by employing an ANCOVA to statically test for the presence of a regression discontinuity effect (Braden & Bryant, 1990; Cook & Cambell, 1979). First, we tested the posttest for homogeneity of regression and the results were found to be non-significant—$F < 1$. Thus, we were able to conclude that there was no difference in slope between the two groups.

According to the results of the ANCOVA, the adjusted mean scores associated with the posttest for participants in the TiPS group ($M = 20.24$, $SE = 1.05$) were statistically significantly higher than those of their peers in control group ($M = 16.47$, $SE = 0.50$), $F(1, 154) = 8.05$, $MSE = 20.65$, $p = .005$. Cohen's d statistic for these data yields an effect size estimate of .46, which corresponds to a medium effect. Overall, the results indicate a positive, practical effect that can be attributed to the TiPS instruction.

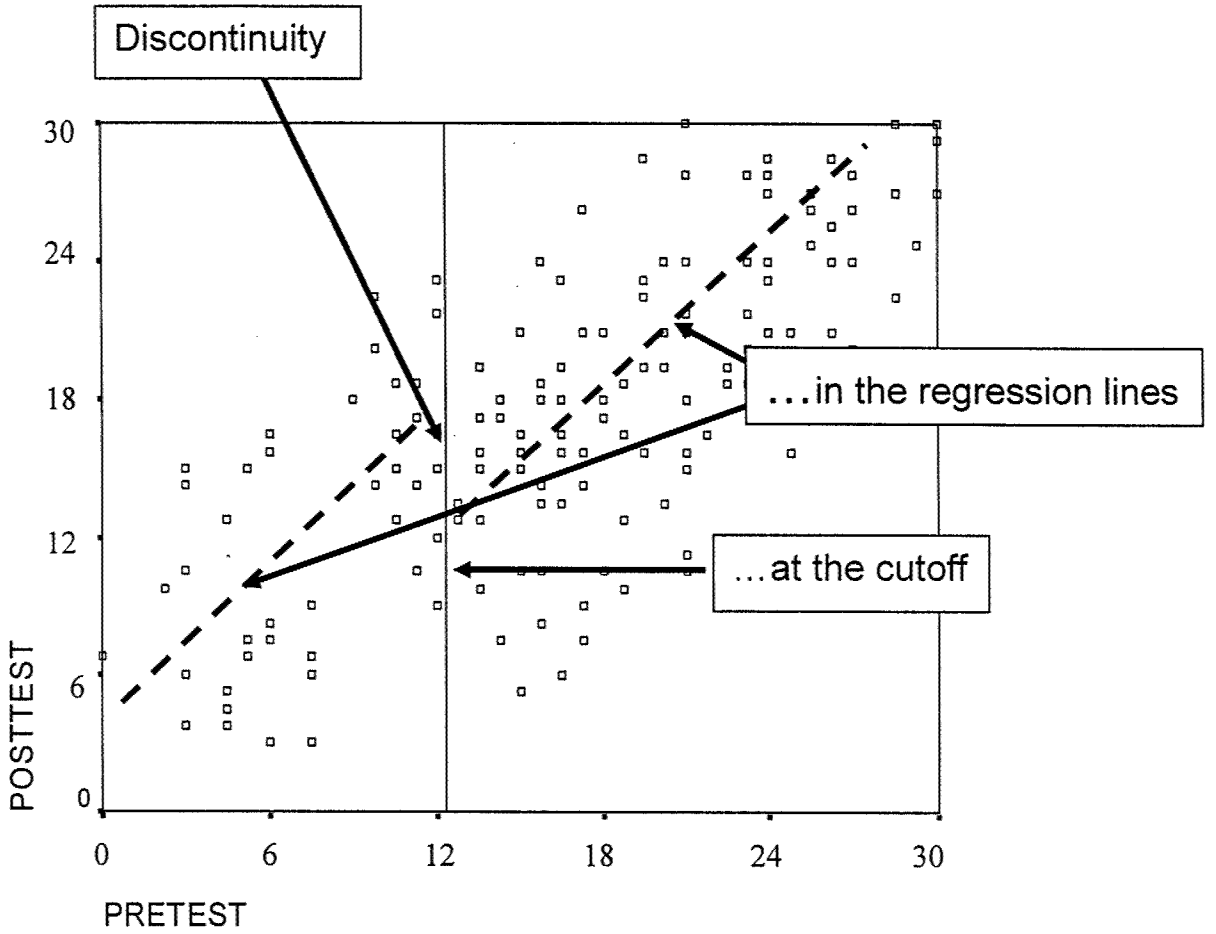


Figure 7. Graph of the pretest scores and posttest scores of the TiPS and control participants. The TiPS group's regression line is represented by the dashed line on the left and the control group's regression line is represented by the dashed line on the right.

As previously mentioned, since it was possible for the ANCOVA to be misspecified (e.g., the shape of the regression surface is not properly modeled due to a curvilinear relationship between the pretest and posttest), we attempted to exactly specify the true model by following the steps outlined in the analysis section. We pursued this model specification process since it we felt that it would help ensure that we would not erroneously conclude the TiPS treatment made a difference when it in fact did not.

First, we regressed the posttest scores on the modified pretest (SPSS variable = "precut"), the treatment variable (SPSS variable = "group"), linear interaction (SPSS variable = "linint"), higher order transformation including quadratic (SPSS variable = "quad") and quadratic interaction (SPSS variable = "quadint"). The result of our initial specifications of the model is described below:

Estimate Initial Model

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	QUADINT, QUAD, GROUP, LININT, PRECUT[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: POSTTEST

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.750[a]	.563	.548	4.56781

a. Predictors: (Constant), QUADINT, QUAD, GROUP, LININT, PRECUT

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4053.462	5	810.692	38.854	.000[a]
	Residual	3150.601	151	20.865		
	Total	7204.063	156			

a. Predictors: (Constant), QUADINT, QUAD, GROUP, LININT, PRECUT

b. Dependent Variable: POSTTEST

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	12.582	1.374		9.158	.000
	PRECUT	.734	.348	.781	2.108	.037
	GROUP	4.351	1.951	.282	2.230	.027
	LININT	.903	.771	.351	1.171	.244
	QUAD	.007	.019	.084	.364	.716
	QUADINT	.068	.071	.225	.951	.343

a. Dependent Variable: POSTTEST

The treatment effect estimate is the one next to the "group" variable. The initial estimate is 4.35 (SE = 1.95)—very close to our estimated treatment effect of 4 derived from our visual examination of Figure 7. However, there was also evidence that several of the higher-order terms

were not statistically significant (see *t*-values on right side of table) and, thus, were not needed in the model. To refine the model, we dropped the two quadratic terms (see below):

Refining the Model

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	LININT, GROUP, PRECUT[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: POSTTEST

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.748[a]	.559	.550	4.55750

a. Predictors: (Constant), LININT, GROUP, PRECUT

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4026.135	3	1342.045	64.612	.000[a]
	Residual	3177.928	153	20.771		
	Total	7204.063	156			

a. Predictors: (Constant), LININT, GROUP, PRECUT

b. Dependent Variable: POSTTEST

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	12.190	.853		14.294	.000
	PRECUT	.856	.088	.912	9.714	.000
	GROUP	3.937	1.445	.255	2.725	.007
	LININT	.068	.225	.027	.303	.762

a. Dependent Variable: POSTTEST

In this refined model, the treatment effect estimate was 3.94 and the *SE* of 1.45, which is lower than the initial model. This indicated a gain in efficiency due to the elimination of the two unneeded quadratic terms. However, again we found evidence that the linear interaction was not statistically significant and, thus, were not needed in the model. To refine the model, we dropped the non-significant linear interaction term and respecified the model (see below).

Final Model

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	GROUP, PRECUT[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: POSTTEST

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.747[a]	.559	.553	4.54404

a. Predictors: (Constant), GROUP, PRECUT

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4024.224	2	2012.112	97.447	.000[a]
	Residual	3179.838	154	20.648		
	Total	7204.063	156			

a. Predictors: (Constant), GROUP, PRECUT

b. Dependent Variable: POSTTEST

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	12.102	.800		15.136	.000
	PRECUT	.867	.081	.923	10.720	.000
	GROUP	3.767	1.328	.244	2.837	.005

a. Dependent Variable: POSTTEST

In the final model, the treatment effect and *SE* were almost identical to the previous model and all of the terms were statistically significant, indicating that this final model fit the data well and, thus, did not need any further refinement. This also indicated that there was no evidence of a curvilinear relationship associated with the bivariate pretest-posttest relationship. Instead, we were able to conclude that a straight-line model, such as the one assumed in our aforementioned ANCOVA analysis, accurately captures this data. As evidence, this model—like the other analysis—indicated that the TiPS treatment produced a statistically significant effect, $t(154) = 2.837$, $p = .005$. In fact, the results of our ANCOVA and our model specification process were identical (in a two group situation, $t^2 = F$; thus, squaring our t-value or 2.837 equals 8.05, the F-value we obtained from our ANCOVA).

Beyond providing a basic evaluation of the overall TiPS system, we wanted to determine the amount of instructional time involved with the use of TiPS system The average time for completing the Tips computer instruction was **3.84 hours** (*SD* = 0.97) and ranged from 1.83 to 6 hours (excluding test-taking time), during which time they completed an average of **64.28 problems** (*SD* = 7.68) on the system (overall range from 54 to 80 problems). This instructional time result diverged from the results of the previous field trial where it was found that the average time for completing the Tips computer instruction was 5.81 hours and ranged from 3.63 to 9.52 hours (excluding test-taking time). However, unlike the field trial that permitted the student to work on TiPS in session ranging from three to six hours, the participants in the present study were not be permitted to work on TiPS for more than an hour a day. This latter type of arrangement may have encouraged the students to use their time more efficiently on the system.

With regard to our final objective, the affective nature of the TiPS learning experience, we examined how the TiPS students responded to the affective questionnaire. In response to the statement:

- "I have learned to solve word problems based on this instruction", 21 out of 41 (51.2%) students agreed or somewhat agreed.
- "Learning was fun", 27 out of 41 (61%) students agreed or somewhat agreed.
- "I would prefer learning from TiPS when I have to study 'mathematized' contents next time", 27 out of 41 (61%) students agreed or somewhat agreed.
- "I felt curious", 23 out of 41 (56.1%) students agreed or somewhat agreed.
- **"The examples and problems in TiPS helped me to understand word problems", 32 out of 41 (78.1%) students agreed or somewhat agreed.**
- "I was interested in learning about word problems" 20 out of 41 (48.8%) students agreed or somewhat agreed.
- **"The instruction in TiPS was well designed", 31 out of 41 (75.7%) students agreed or somewhat agreed.**

Comparing TiPS Group to Eligible/Untreated Group

To examine the performance of the TiPS group relative to the eligible/no treatment condition, pretest to posttest gain scores were calculated for these two groups and analyzed with an independent sample *t*-test. According to the results of the *t*-test, there was a statistically significant difference between the posttest performance of the participants in the TiPS condition and their peers in the eligible/no treatment, $t(80) = 2.23$, $p = .029$. The participants assigned to the Cohen's *d* statistic for these data yields an effect size estimate of .50, which corresponds to a medium effect. Again, this result indicates a positive, practical effect that can be attributed to the TiPS instruction, as opposed to some intervening relevant instructional experiences (inside or outside of the study) that perhaps all of the students at Mississippi State that performed below the cutoff criterion were exposed to during the course of this study.

It is also worth noting that, according to the results of an ANCOVA, the adjusted mean scores associated with the **posttest** for participants in the **eligible/untreated group** ($M = 18.09$, $SE = 0.97$) were **not statistically different** than those of their peers in **untreated control group** ($M =$

16.47, $SE = 0.50$), $F(1, 154) = 8.05$, $MSE = 20.65$, $p = .29$. Moreover, our attempts to specify an analytic regression model in this case did not produce statistically significant results (see below).

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	12.582	1.306		9.638	.000
	GROUP	1.791	1.934	.116	.926	.356
	PRECUT	.734	.331	.738	2.219	.028
	LININT	1.062	.909	.336	1.169	.244
	QUAD	.007	.018	.084	.383	.702
	QUADINT	.124	.095	.296	1.304	.194

a. Dependent Variable: POSTTEST

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	12.190	.813		15.003	.000
	GROUP	.876	1.352	.057	.648	.518
	PRECUT	.856	.084	.861	10.196	.000
	LININT	-.199	.250	-.063	-.794	.429

a. Dependent Variable: POSTTEST

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	12.378	.776		15.946	.000
	GROUP	1.316	1.231	.085	1.069	.287
	PRECUT	.834	.079	.839	10.555	.000

a. Dependent Variable: POSTTEST

Taken together, this implies that none of the 124 students that scored below the cutoff criterion would have been able to produce (statistically) significantly higher posttest scores, after adjusting for pretest performance, than their peers in the untreated control group without the targeted intervention provided by TiPS.

Conclusions

In sum, it is apparent from the evidence compiled in during the present project that remedial learners engaged in mathematical thinking can benefit on a variety of cognitive (i.e., transfer)

and affective measures by working within TiPS, a computer-based learning environment designed to develop learners' problem solving skills. In particular, we empirically documented that learners excelled—after who spent an average of four hours on the TiPS system were typically rewarded with a 15% improvement—a 1½ letter grade improvement, by conventional standards—in their problem-solving performance. We attribute this effect to the collection of features inherent to TiPS, including: (a) instruction within the context of problem solving scenarios, designed to gradually build the learners skills and abilities that should enable them to reason about complex real-world problems, (b) auxiliary representations depicted in its problem-solving interface to help learners model and solve problem situations, (c) worked examples to provide the learners with an expert's solution, which they can use as a model for their own problem solving, and (d) the Bayesian student modeler that monitors the learners performance and adapts instructional delivery to meet their needs. In addition to the enhanced problem-solving performance, the learners exposed to TiPS reported feeling that the instructional material, including the examples and problems, helped them to understand how to approach and solve word problems and that, overall, the instructional environment was well designed.

REFERENCES

Atkinson, R. K., & Derry, S. J. (2000). Computer-based examples designed to encourage optimal example processing: A study examining the impact of sequentially presented, subgoal-oriented worked examples. In B. Fishman & S. F. O'Connor-Divelbiss (Eds.), *Proceedings of the Fourth International Conference of Learning Sciences* (pp. 132-133). Hillsdale, NJ: Erlbaum.

Atkinson, R., Derry, S. J., Renkl, A. & Wortham, D. (2001). Learning from examples: Instructional principles from the worked examples research. *Review of Educational Research, 70,* 181-215

Atkinson, R. K., Wortham, D. W., Derry, S. J., Jiang, N., & Gance, S. K. (1998). *Beyond static worked examples: Testing the efficacy of computer delivered dynamic examples.* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Barlow, D. H., & Hersen, M. (1984). *Single case experimental designs: Strategies for studying behavior change* (2nd ed.). Boston, MA: Allyn and Bacon.

Braden, J., P., & Bryant, T. J. (1990). Regression discontinuity designs: Application for school psychologists. *School Psychology Review, 19,* 232-240.

Campbell, D. T., & Cook, J. C. (1963). *Experimental and quasi-experimental designs for research.* Boston, MA: Houghton Mifflin.

Carpenter, T.P. & Moser, J.M.(1982). The development of addition and subtraction problem-solving skills. In Carpenter, Thomas P., Moser, James M., and Romberg, Thomas A. (Eds.) *Addition and Subtraction: A Cognitive Perspective.* Hillsdale, NJ: Erlbaum.

Catrambone, R. (1994). Improving examples to improve transfer to novel problems. *Memory & Cognition, 22,* 606-615.

Derry, S. J., Tookey, K., Smith, C., Potts, M. K., Wortham, D. W., & Michailidi, A. (1994). *Psychological foundations of the TiPS system: A handbook for system 1.0* (Technical Report). Madison, WI: Wisconsin Center for Education Research.

Derry, S., Weaver, G., Liou, Y., Barker, J., & Salazar, E. (1991). *Inducing flexible transfer in novice problem solvers: Effects of three instructional approaches.* Unpublished manuscript.

Derry, S. J., Wortham, D., Webb, D., & Jiang, N. (1996). *Tutorials in problem solving (TiPS): Toward a schema-based approach to instructional tool design.* Paper presented at the annual meeting of the American Educational Research Association, New York, NY.

Gitomer, D. H., Steinberg, L. S., & Mislevy, R. J. (1995). Diagnostic assessment of troubleshooting skill in an intelligent tutoring system. In P. Nichols, S. Chipman, & S. Brennan (Eds.), *Cognitively Diagnostic Assessment*. Hillsdale, NJ: Erlbaum.

Larkin, J., McDermott, J., Simon, D. and Simon, J. (1980). Models of Competence in solving physics problems. *Cognitive Science, 4,* 317-345.

Levin, J. R., Serlin, R. C., & Seaman, M. A. (1994). A controlled, powerful multiple-comparison strategy for several situations. *Psychological Bulletin, 115,* 153-159.

Marshall, S. P. (1995). *Schemas in Problem Solving*. Cambridge, England: Cambridge Press.

Quilici, J. L., & Mayer, R. E. (1996). Role of examples in how students learn to categorize statistics word problems. Journal of Educational Psychology, 88, 144-161.

Renkl, A. (1997). Learning from worked-out examples: A study on individual differences. *Cognitive Science, 21,* 1-29.

Riley, M. S., Greeno, J.G. & Heller, J.I. (1983). Development of children's problem-solving ability in arithmetic. In H. P. Ginsberg (Ed.), *The development of mathematical thinking* (pp. 153-196). Orlando, FL: Academic.

Seaman, M. A., Levin, J. R., & Serlin, R. C. (1991). New developments in pairwise multiple comparisons: Some powerful and practicable procedures.

SPSS-X. (1998). *SPSS-X User's Guide* (3rd Ed.). Chicago, IL: SPSS Inc.

Sweller, J. (1993). Some cognitive processes and their consequences for the organisation and presentation of information. *Australian Journal of Psychology, 45,* 1-8.

Tookey, K. R., & Derry, S. J. (1994). *Arithmetic schema and strategy instruction: A study of two problem-solving approaches.* (ONR Technical Report 94-2) Madison, WI: Wisconsin Center for Education Research.

Trochim, W. (2001). *Research methods knowledge base*. Cincinnati, OH: Atomic Dog Publishing.

VanLehn. K. (1996). Cognitive skill acquisition. In J. Spence, J. Darly, & D. J. Foss (Eds.), *Annual review of psychology*: Vol. 47. Palo Alto, CA: Annual Reviews.

Van Merrienboer, J. G., & De Crook, M. B. (1992). Strategies for computer-based programming instruction: Program completion vs. Program generation. *Journal of Educational Computing Research, 8,* 365-394.

Ward, M., & Sweller, J. (1990). Structuring effective worked examples. *Cognition and Instruction, 7,* 1-39.

Wortham, D. W. (1996). *Testing a learning environment based on a semantic analysis of math problems.* Master's Thesis, Department of Educational Psychology, University of Wisconsin-Madison.

Wortham, D. W., Webb, D. C., & Atkinson, R. K. (1997). *Effects of two curricula on proportional reasoning, problem solving, and strategy use.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

TABLE 1

Unadjusted Mean Performance on Pretest, Posttest, and Gain Scores for Phase 2 by Condition.

	MEASURE					
	Pretest		Posttest		Gain Score	
Condition	***M***	***SD***	***M***	***SD***	***M***	***SD***
Untreated Control Group	20.40	4.82	19.38	6.06	-1.02	4.49
TiPS Group	7.54	3.48	12.00	5.76	4.46	4.78
Eligible/Untreated Group	8.35	3.30	10.72	4.37	2.29	4.05

APPENDIX A

Anecdotal Comments/Observations

Recently, I asked several graduate students enrolled in my Intelligent Tutoring System course at Arizona State University to work with TiPS while trying to behave as they would expect a low-ability student to interact with the system and then to provide me with feedback regarding the experience. Their anecdotal comments/observations are described in the following section.

Graduate Student 1

Experience

As requested, I entered the system and floundered through it. Even without directions, it was fairly intuitive. I was able to navigate between the problem sets and the problems themselves. At the problem set level, it was fairly easy to select the categories I wanted to navigate. Since I wasn't sure whether there was a specific order, I jumped around. This presented no learning problems, as the units appeared to be atomic. This was fine, until I attempted one of the transfer problems, where the user is required to select the correct model. Because I had not proceeded in a linear order through the instructional problems, I scored poorly.

Within the problem sets, I thought the instructional animation presenting the format of each model was good. Although I was not sure why I was applying this in every case, I know exactly what to do. When I made a mistake, I had a hard time erasing incorrect portions of the problem without clearing the entire problem. The help was very nice, but followed a very dramatic curve. Minor assistance was given, then the whole solution was presented. While I was able to then correctly complete the problem, I didn't learn much. For problem 2A of the vary lesson, I copied the correct model, but was only able to get 9 out of 10. I thought the scoring was appropriate. As was the feedback with the scoring. All the practice questions appeared to be of equal complexity. When checking my mastery, I had a hard time detecting a change from one reading to the next. If they moved, the difference was slight and often I could not remember the previous setting.

In getting a 10/10 mastery, I did notice that certain practices were grayed out and that the mastery score appeared to be higher.

Feedback

Most of my issues deal with the user interface. I thought it was cumbersome and unevenly spaced. Having two different prompt pop-ups was jarring. Rather than having consistent prompts, some looked like regular operating system prompts. On the first one, I thought the operating system was having a problem. The buttons were not grouped in a normal pattern. I thought the helps were very common, and progressed to the answer too quickly. It would have been nice if each question became progressively more difficult and the hints less instructive. This way, students would not be able to get the answer every time by running through hints. I

would also like to know the weighting of each hint. I could not tell if I was penalized more for the third help or if I was equally penalized, even though the last help was much more valuable than the first. I would have also like a bit more instruction with each model and maybe instruction in the differences between all four at the beginning. Rather than letting the student pick any module, they should be instructed to proceed in a linear fashion, or only un-gray the buttons that they can select at random. For the mastery section, I would like to have lines across the bar graph. That way I could track my incremental progress from one problem to the next. At the end when I tried to exit, I received a schema failure. It dumped all of my data. I was running it on Windows XP. Finally, I would like to know when I was finished with the program.

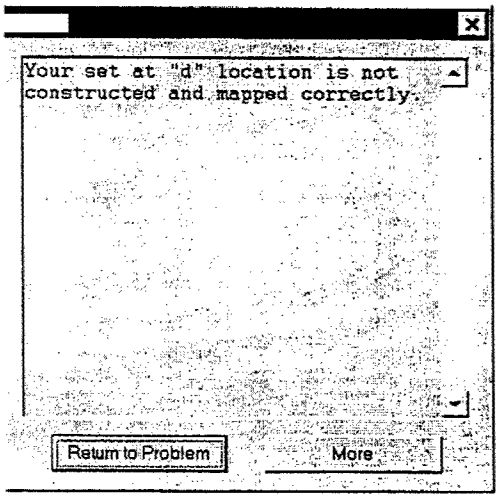
Graduate Student 2

1. Installed on 3/1/03, without any difficulty, under Windows 98SE
2. The bottom of the page was underneath the task bar on a 800x600 resolution monitor, requiring me to move the task bar.
3. First impression – if the Navy personnel who are to use this have math skills that are that bad, chances are their reading skills will not be up to the level required by these word problems.
4. Is this the "new new math"? I have never seen this method of "instruction by diagramming" for these skills. I found it somewhat confusing.
5. The tutorial should first briefly address "what are these shapes in the buttons", and reassure student that they will be fully explained later.
6. The tutorial starts by demonstrating an error, without first informing the student that this is intentional. A student, who actually understands the process, may get confused at this point.
7. There was no feedback to inform me when I had finished ("jumped through all the right hoops"), until I pressed "Done" or "Hint". I pressed "Hint" most often, since I was not always sure I was really done.
8. If I choose the "wrong" tool, but completed the problem, I got 1/10 points.
9. I disliked the insistence on putting all the "labels" in the right spot, in addition to the units. This is not how I would work the problem.

Graduate Student 3

1. I noticed the change in the Mastery Levels as 1 fluctuated how well I was doing but I didn't notice any change in the way TIPS operated or if it adjusted the level of difficulty.
2. It was easy to be perfect when I tried, but one thing bothered me was that I didn't see an increase in Master Scores until I did a lot. It would have been nice to have the Mastery Scores show: current level, increase, and total (maybe using different colors).
3. Some of the icons were confusing, but since the same one was used for each problem in a section, once you had the right one you just kept choosing it until the section was done. This made choosing the right tool kind of pointless.
4. Also, the tutoring said to double-click (if I remember correctly) while it was really necessary to "highlight" the term to be added and then adding it. For example, "non-produce" can't be selected by double-clicking (you can select "non" or "produce", but not both).

5. The Hints were not very helpful (see image), and even if I used the Hints three times for each step in a problem, I still got 10/10! The Hint references "d" location. What "d"? None of the tools had a "d" section that I could find.



6. In the variable tools section, different tools could be used to answer the same problem. When I used a grouping tool instead of the change tool, I got 1/10 even though I had the correct answer! The feedback would have been more helpful if it had suggested solving the problem would have been easier with an alternate tool.
7. It is difficult to gauge what grade-level this is designed for. The problems seem around middle school, but learning how to use the interface and do the problems seem more high school range.
8. The nature of the problems is nicely varied and having the stories to go along with the word problems from item to item made it more interesting (I really felt like I got to know how bad a driver Lucy is.
9. The "sound effect" during the tutorial was distracting. I had to go through it twice before I really had an understanding of how to work the problems.
10. Of course the layout, visual appeal, etc. need to be worked on (but since it is functional, I wouldn't worry too much about the cosmetics of it).

Graduate Student 4

My main points of criticism towards tips are the following:

- The problems are often wordy and lengthy and even if this is mirroring true conversation I believe it has negative implications on the students' abilities to learn the intended methods. The problem should be stated and the small-talk left aside.
- The interface of highlighting and moving words, numbers etc is cumbersome. It would have been much better if the student could just simply click on the (already colored) word and then click in the method diagram. It was easy to miss and that made it frustrating.
- The mastery graph does not give very much information. It is either missing labels or there is something wrong with the software. As it stands now it doesn't give much feedback at all.

- The hint system is neat but a little stodgy (?). It always start with general comments like "It is important to build correct sets." That to me is almost like an insulting comment because it is so obvious. However, the hints further into the solution are more adequate and detailed.
- The evaluation often criticizes perfect solutions because they don't have the exact same wording. For instance: when the set said "dealer wins" tips evaluation wanted to see "won by the dealer".

Overall, it seems like tips is well tested and worked through but it has interface imperfections by 2003 standards and that could be accepted since it is ten years since the first version came out and five years since the last major revision.

Suggestions for Enhancing System

One of the graduate students that worked closely with TiPS at Mississippi State was a former secondary math teacher that taught algebra and pre-algebra to students "at-risk" in their math learning skills. The following section contains his suggestions as to how one might enhance the system (beyond fixing the issues raised in the previous sections):

Change Tool Problems

It may be beneficial to precede this module with some instruction on the various words that indicate positive change versus those that indicate negative change. The students will then have a better idea of when to solve problems by adding the amounts and when to solve them by finding the difference. (Examples: "taken" vs. "given", "payment" vs. "borrowed", "credit" vs. "debit", "increased" vs. "decreased", "advance" vs. "drop", etc.)

Choice of Tool/Scoring

It is usually helpful to distinguish between problems that deal with changes, comparisons or groupings, but there can be cases where the same problem could be interpreted by different students to fall under any of these categories. As long as the student makes sense of the problem and solves it logically, it is usually best to allow full credit on the scoring. An example follows:

> The two longest railway tunnels in the world, the Seikan and the Dai-shimizu, are both in Japan. The Dai-shimizu tunnel is 14 miles { mi } long. The Seikan tunnel is 19.5 miles { mi } longer. How long is the Seikan railway tunnel?

A student using the change tool may interpret this problem as follows: "If 19.5 miles of tunnel were added to the 14-mile Dai-shimizu tunnel, it would be as long as the Seikan tunnel. How long is the Seikan tunnel?"

A student using the compare tool may interpret this problem as follows: "If the Dai-shimizu tunnel and the Seikan tunnel are compared, the length of the Seikan tunnel is 19.5 miles greater in length. The Dai-shimizu tunnel is 14 miles long. How long is the Seikan tunnel?"

A student using the grouping tool may interpret this problem as follows: "If a 19.5 mile tunnel is combined with the length of the 14-mile Dai-shimizu tunnel, the total would be the length of the Seikan tunnel. How long is the Seikan tunnel?"

The Function Tool

The function tool model appeared a bit confusing. Simply having an arrow pointing from the X box to the "func" box to the Y box was not helpful in setting up or solving the problems. All of the problems in this module involved only multiplying a constant times X in order to get Y. A clearer model, in this case, might have simply been as follows: Y or f(X) = kX where k is a constant value.

Vary Lessons (Proportions)

The word "proportion" seems to be more descriptive than the word "vary". Also, when describing a proportion, the format of "A is to B the same way that C is to D" is usually less confusing to students than using a Vary format of "If A = B, then C = D." The reason why the vary format may confuse some is that it leads to false statements like "If 5% = 100%" or "If 3 plums = 2 apples" which everyone knows is untrue. A better format would be, for example, "5% is to 100% the same way that $2.50 is to what amount?"

Student performance on Vary lessons might also be improved by furnishing some additional guidance at the beginning. It might be helpful, for example, to precede the lessons with clear instructions that proportions must be set up with the same units being compared.

APPENDIX B

Items from Pre-and Posttests Used in Phase 2

Problem type	Form A	Form B
Function	1. Jerry sold 2.5 times more cups of lemonade yesterday than he sold today. If Jerry sold 18 cups of lemonade today, how many cups did Jerry sell yesterday?	1. Tom was 1.5 times faster than Jeff in reviewing his math test. If Jeff spent 3 hours to review his math test, how long would Tom spend?
Vary	2. If you want to go rafting on the Dolores River in Utah, an outfitter will rent you rafts that can hold 6 passengers each. How many rafts will you need if there are 42 passengers in your group?	2. Helen will be traveling in a car, which averages 25 miles per gallon of gas. If she wants to travel 350 miles, how many gallons of gas does Helen need?
Vary	3. A diagram measuring 16 cm long is reduced on a copying machine to 10 cm long. If the width of the original diagram is 12 cm, what is the width of the reduced copy?	3. Jimmy and his brother made a model of their tree house. In the real tree house, one of the walls is 27 inches wide by 63 inches long. The corresponding wall in the model is 32 inches long. How wide is this wall in the model?

Group and Vary	4. Pam took 72 pictures on her trip. If she can take 4 good pictures for every 5 bad pictures, how many bad pictures did she take on her trip?	4. To make a certain amount of plate glass, every 18 kilograms of sand are combined with 7 kilograms of other ingredients. How many kilograms of sand are needed to make 100 kilograms of glass?
Vary *NAEP problem and NAEP isomorphic problem*	5. John makes lemonade using 15 spoonfuls of sugar and 21 spoonfuls of lemon. Mary makes lemonade using 11 spoonfuls of sugar and 17 spoonfuls of lemon. Which recipe will taste sweeter or will they taste the same? Give mathematical evidence to justify your answer. ***(NAEP isomorphic problem)***	5. Luis mixed 6 ounces of cherry syrup with 53 ounces of water to make a cherry-syrup. Martin mixed 5 ounces of the same cherry syrup with 42 ounces of water. Who made the drink with the stronger cherry flavor? Give mathematical evidence to justify your answer. *(NAEP Problem)*
Function and Vary	6. Tom wants to take the new train from Chicago to New York, a distance of 790 miles. The new train is 1.5 times faster than the old one that could travel at 120 mph. How long will it take Tom to get to New York on the new train?	6. The textile factory has an old machine that is able to spray a waterproofing solution on 24 square yards of fabric in 1 hour. The factory just installed a new machine that is 2.5 times faster than the old one. How long will the new machine take to spray 150 square yards of fabric?

Vary and Group *NAEP problem and NAEP isomorphic problem*	7. Bill purchased a new car that was selling for a price of $19,200. He paid $4,800 as a down payment and obtained a 36-month car loan to finance the reminder of the selling price. If his monthly payment were $451, what is the total amount, including the down payment, that Bill will pay for the car? *(NAEP problem)*	7. Allen will buy a used boat that is advertised for a price of $21,600. He will pay $ 3,600 as a down payment and obtain a 48-month loan to finance the reminder of the selling price. If his monthly loan payment is $442, what will be the total amount including the down payment, that Allen pays for the boat? *(NAEP isomorphic problem)*
Function, Vary, and Vary	8. Martin is traveling to visit friends in a nearby town. He plans to take the bus to their house, and then to ride his bike home after the visit. On his bike ride Martin will travel the same distance as he traveled by bus. It takes Martin 4.6 minutes to go one mile on his bike, which is 4 times faster than it takes the bus to travel one mile. If it takes the bus 138 minutes to reach his friends' house, how long will it take Martin to make the round trip? (*Note: 60 minutes = 1 hour*)	8. Beverly called Ace Moving Company to get an estimate of the amount of time it would take them to pack her house. Ace told her it takes 68 min to pack a kitchen box, which is 4 times longer than it takes to pack a box from any other room. They will use the same number of boxes for the kitchen as will be used for all the other rooms of house together. If it takes the movers 510 minutes (8.5 hours) to pack the kitchen, how long will it take to pack Beverly's whole house?

Vary, Change, and Vary	9. Jeff has a pile of 250 logs that must be split. He and four friends were able to split 100 of them in two and a half hours on Monday. The following day, he and two of his friends decided to finish the job. How long did it take the three of them to complete splitting all the logs?	9. Henry and his friends have 85 fish to clean. On Friday, he and three friends cleaned 20 of the fish in one and a half hours. Saturday, only two friends returned to help clean. How many hours will I take the three of them to finish cleaning all of the fish?
Group, Vary, and Vary	10. Rosa's company reimburses her for mileage she puts on her car while doing business. Last month, she made three trips in the United States of 45, 60, and 35 miles, and received $150. This month she drove to Mexico, but she forgot to record her car's odometer reading when she started the trip. According to the road signs in Mexico, however, she traveled 143 kilometers one way, or 286 kilometers round trip. How much will she receive for the Mexico trip? (1 km = .625 miles)	10. Jared's club sells lemonade to make money during performances at the campus theater. Last week, they made enough lemonade to fill three large thermos jugs. One jug held 18 quarts, another held 20 quarts, and the third held 25 quarts. They sold all the lemonade and made $120. This week, they plan to use one large thermos that will hold 50 liters. If they sell all 50 liters, how much will they make? (1 quart = .946 liters).

APPENDIX C

Guidelines for Deriving Conceptual Score on Pretest and Posttest Items

Points Assigned	Scoring Guidelines
3 points	Correct (ignoring minor computational/copying errors) • There is perfect understanding of the problem and the student used a complete and correct strategy to arrive at an answer. The only error allowed is a simple calculation error (such as 3 X 4 = 13)
2 points	Substantial understanding of the problem • The student basically understands the problem and is pursuing an identifiable solution strategy that is essentially correct. However, the student has made one or two fatal errors, such as mislabeling problem facts, confusing arithmetic operations, leaving out a small step, etc. On this type of error, a tutor probably would not have to intervene with full explanation or make a student start over, but probably would help the student identify and correct error(s) in the current path.
1 point	Low understanding of the problem. • There is at least some slight evidence that one or more concepts underlying the problem are understood, although conceptual understanding of the problem is significantly flawed. In tutoring, a significant amount of explaining and coaching would be required in order to repair the problem solution, and the student would most likely have to start over in order to work the problem correctly (as opposed to making a repair in the current solution strategy).
0 points	No understanding exhibited. • There is no evidence of the student's understanding the problem.